Advantages and limitations of Twin Assessment of Clinical Trials (TACT)

Porzsolt, Franz; Costa, Ian Curi Bonotto de Oliveira; Thomaz, Tania Gouvêa

J Public Health (2009) 17:425–432
DOI 10.1007/s10389-009-0283-4

ORIGINAL ARTICLE

Section Original article

Advantages and limitations of Twin Assessment of Clinical Trials (TACT)

Franz Porzsolt · Ian Curi Bonotto de Oliveira Costa · Tania Gouvêa Thomaz

Received: 9 February 2009 / Accepted: 15 July 2009 / Published online: 26 August 2009

Abstract
Background Considerable time and energy are expended in the scientific community to discuss the validity, importance, and applicability of the results of clinical trials. Depending on the goals, perspectives, and other motivating factors, protagonists and skeptics come to different conclusions, even when using the same methods and tools for critical appraisal. The aim of this study was to complement existing methods and tools with minor modifications to provide a prototype instrument that generates commonly accepted versions of critical appraisals.
Methods As a pilot experiment, one university-based and one industry-based referee independently completed the twin assessment of five trials published in well-recognized journals. They identified the study questions, defined the simplest, i.e., ideal, study designs to answer these questions, and checked eight validity criteria. Identical positive or negative answers of both referees increased or decreased the validity score. A maximum of two disagreements (0 score) was allowed. This procedure, which had been tested by two referees in a pilot experiment, was repeated with 19 third-year medical students and their supervisor at the Universidade Federal Fluminense, Niterói/RJ, Brasil. Four students each played the roles of the industry-based and university-based referees and finally recorded their consensus.
Results The two referees of the pilot experiment agreed in all but one answer to the five investigated publications. The points of criticism differed in various papers. The consensus reached by the students considerably differed from the consensus reached by the referees.
Conclusions A consensus score generated by two referees or by two groups of students is feasible, but the achieved result is not necessarily reproducible. The critical appraisal of the study question in connection with the applied study design deserves special attention. It is time consuming but possible to identify and describe the possible flaws in the design conduct and report of clinical trials, but it is unlikely to reach a reproducible interpretation. These data indicate the problems with even evidence-based assessments and appraisals: the assessments may well be reproducible, but not the appraisals. Quality scores that include also the appraisal may therefore be interpreted with caution. Appraisals or quality scores may be used for interim decisions until data are available that confirm under real-world conditions what was predicted by the results generated under ideal but artificial conditions of a clinical trial.

Keywords Assessment · Critical appraisal · Validity · Consensus · Confirmation-based health care

Members of SHUFFLE (Students Help Universidade Fedeal Fluminense in Literature Export) are Prof. Dr.Tania G. Thomaz, MD, Ian C.B.O. Costa (manager), Priscila F. de Amorim, Rafael C. Astorga, Renato C. Branco, Igor Brandão, Gustavo M. Cardoso, Lara Dan, Carolina D.Gonçalves, Thiago Gonçalves, Tabata Graciolli, Davi L.Grilo, Ingrid E. Hinden, Pedro M.Lisboa, Gabriela M. Machado, Pascale G. Massena, Lauanny A. Pereira, Rafael Perez, Natalia C.Z. Silva, and Fábio J.S. Souza.

F. Porzsolt (✉)
Clinical Economics, University of Ulm,
Frauensteige 6,
89075 Ulm, Germany
e-mail: franz.porzsolt@uniklinik-ulm.de

I. C. B. de O. Costa · T. G. Thomaz
SHUFFLE, Dept Physiology, Federal University Fluminense,
Niterói, Rio de Janeiro, Brasil

Introduction

Jørgensen et al. (2006) are probably correct in their conclusions in a recent article that "industry-supported

reviews of drugs should be read with caution, as they were less transparent, had few reservations about methodological limitations of the included trials, and had more favorable conclusions than the corresponding Cochrane reviews" (Jørgensen 2006). We are not sure that colleagues who work in fundamental research and patient care—and write the above-mentioned articles—behave with greater integrity than colleagues in industry, but we are confident that they have different interests and are motivated by different incentives. The real problem is that these scientists and practicing physicians in the public sector and those engaged in industry are not the only stakeholders in the health-care system. Hospital managers, network managers, and health insurers also have slightly different perspectives and interests. Many people realize the problem, but are reluctant to articulate it. Therefore, Jørgensen's statement has to be supplemented by a second statement indicating that publications are one way of 'opinion framing' influencing opinions, but there are many other ways to influence health-care decisions.

Falsehoods may result from corrupted evidence or corrupted dissemination of otherwise valid evidence. These falsehoods, when consumed as truth by unwitting and well-intentioned practitioners of EBM, then disseminated and adopted as routine practice, may well result not only in inappropriate quality standards and processes in health care, but also in harm to patients. For example, the practicing radiologist today knows that earlier diagnosis of lung cancer can be achieved. The medical practitioner also is confident that intervention after early diagnosis is more effective in preventing death in this otherwise fatal disease. The practitioner is thus inclined to consider such screening in a high-risk person with suitably long life expectancy, especially when asked to provide it. Official recommendations against lung-cancer screening, said to be based on demonstrated lack of effectiveness of traditional radiographic screening, create a dilemma for the critically thinking practitioner who suspects that something may be seriously wrong. What constitutes evidence? Is it official recommendations or the practitioner's practical experience? Miettinen (Miettinen 2001) finally concludes, "Tomorrow's radiologists will need to be critical thinkers, learning how to read books and journals and to listen to 'experts' more skeptically". This opinion was confirmed by the actual discussion about screening for prostate cancer (Barry 2009). There is probably no other way to acquire personal competence and comprehend what is presented in the scientific literature. It is necessary to understand that informed consent means sharing uncertainty with patients; type-I error means avoiding false optimism, and the intention-to-treat principle is nothing other than a pragmatic analysis (van Gikn 1999). Whether intended or not, omissions and misinterpretations are not conducive to good medical practice and can harm patients and produce an unnecessary financial burden on the health-care system.

Although there are about 40 scoring tools to assess the quality of health information, there is no commonly accepted standard tool (Darmoni 2001) and the validity of checklists is rather low (Forestier 2005). One of the best elaborated instruments is that of the Oxford Center of Evidence-Based Medicine (Phillips). We compared this instrument with even more complex instruments used by the National Institute of Clinical Excellence (NICE) and the German Institute (DIMDI) for appraising health technology assessment (HTA) reports. Our study made two important observations. Despite measuring different things, both instruments recognized the serious weakness of the appraised key publications. Second, despite the weakness mentioned in the core parts of the HTA reports, which are read by scientists, the authors of both reports recommended the use of the procedure in the summaries of their HTA reports, which are read by policy makers (Porzsolt 2005). This asymmetric distribution of information generates problems.

Other examples demonstrate that even Cochrane reviews are not free of weaknesses. We critically appraised the quality of homeopathic studies that were included in a Cochrane review and concluded that the quality of about half of the included studies deserved the lowest validity score of 5 or 6 on a scale of 1-6 (Nothardt 2007). We were also not satisfied with the conclusion in our Cochrane review on the effect of interferon in advanced renal-cell carcinoma (Coppin 2003) and published a different interpretation in the same book together with the same co-authors, but different auspices (Porzsolt et al., 2003a, b). These three little stories demonstrate that two scientists who use the same instrument for assessment of the same study will not necessarily come to the same conclusion. As a consequence there is an urgent need to develop a standard method for evaluating scientific evidence that can detect most of the common biases (Darmoni 2001). In this paper we test an instrument designed to help resolve reviewers' disagreements.

Method

The scoring system

According to the EBM working group (Straus 2005) the assessment of validity of a paper starts with the identification of the investigated scientific question.

At this stage of assessment, we request to define the most simple, i.e., the ideal study design that will answer the study question. Second, this ideal study design is compared with the actually applied design in the investigated study. Third, two referees (one expected pro-advocate and one expected contra-advocate) have to complete independently the twin assessment by answering the same questions and subsequently try to form a consensus.

If both referees agree with their answers one point will be added to the score for each positive answer but one point will be deducted for each negative answer. If the answers of the two referees differ, scoring points will neither be added nor deducted (zero points). As the uncertainty of the assessment increases with increasing numbers of unclear answers, we accept only results up to two disagreements. If the two referees disagree in more than two answers, a third independent referee should be included. As the assessment includes eight questions (Table 1), a maximum of eight points can be added or deducted from the initial score. To avoid negative scores or a score of zero points, we start with an initial score of nine points.

The interpretation of scores

Based on the described rule a maximum of 17 points and a minimum of 1 point can be reached. Table 2 shows all possible combinations. The gray area in Table 2 indicates that this result is acceptable without including a third referee. The validity of a paper can be then classified in five categories: scores of 17 to 15 points indicate optimal validity, scores of 14 to 12 points indicate good validity, scores of 11 to 9 points indicate fair validity, scores of 8 to 6 points indicate borderline validity and fewer than 6 points indicate insufficient validity. For example, assuming two referees agree with a positive answer to four questions, with a negative answer to three questions, and disagree with the answer to one question, the total score will be 10. A score of 10 points classifies the validity of the assessed paper as 'fair.'

Table 1 Scoring system for the Twin Assessment of Clinical Trials (TACT)

No.	Question	Score[§]
-02	Compare ideal with actually applied study design	–
-01	Classify as 'potentially valid'/'not valid' if the 2 designs are likely to produce similar / different results*	Potentially valid
00	If potentially valid, start out with a score of 9 points	09
01	Add/deduct 1 point if the two study designs are similar/different	
02	Add/deduct 1 point if the risk profiles** of the compared study populations are similar/different	
03	Add/deduct 1 point if the allocation of patients to the study groups was concealed/known to the therapist	
04	Add/deduct 1 point if doctors and patients were continuously blinded/were not or not continuously blinded during the entire duration of the study	
05	Add/deduct 1 point if the follow-up was long enough/too short to detect the study endpoints in most of the included patients	
06	Add/deduct 1 point if all/not all included patients were included in the calculation of the study results	
07	Add/deduct 1 point if all/not all included patients were evaluated according to the intent-to-treat principle	
08	Add/deduct 1 point if adequate/inadequate statistics were applied	
	Total score	

[§] A score of "0" indicates that there was no agreement among the two referees in this point. *To classify a study design as potentially valid, it has to be similar to the 'ideal study design.' Otherwise, the reasons for not applying the ideal study design have to be presented. **If one of the study groups has most of the described risks, even if the difference is only minimal, the groups are not really comparable

Reproducibility of the scores

In order to test the reproducibility of the scores, the critical appraisal completed by an industry-based and a university-based referee was repeated by 20 medical students. In order to reduce the workload for the students, each student had to read two papers, one playing the role of a pro-advocate and the other playing the role of the contra-advocate. Four students each played the role of the pro- and four the role of the contra-advocate. These eight students tried to find a consensus applying the above-described instrument and rules to each of the five analyzed papers.

Results

Validation of the chosen scoring system

To start with the validation process of the chosen scoring system, we used papers that were proposed by the initiators of previous EBHC courses. The EBHC students had to rate the validity of these publications which are used to support clinical guidelines. The five included studies were published in Cancer (Dautzenberg 1999), The Lancet (Mehta 2001), The Lancet (Heart protection Study Collaborative Group 2002), New England Journal of Medicine (Meunier 2004 and Piccard-Gebhardt 2005). Initially we planned to use the published articles to demonstrate the difficulties of bias detection without disclosing the publications. As a result of a controversial and critical discussion, we felt we had to disclose the full references and to offer everyone the possibility to respond to our appraisal.

All of these publications described randomized controlled trials and were probably rated as level-I evidence according to the Oxford Centre for Evidence-based Medicine Levels of Evidence (Phillips). The results of our evaluations provided by the referees and by the medical students are shown in Table 3.

The two referees, a pharmacist employed by a pharmaceutical company and a internist (F.P.) working as clinical

Table 2 Score points as functions of the numbers of positive and negative agreements between referees. The grey fields indicate results of assessments in which both referees did not disagree in more than two answers. The validity can be classified according to the number of score points in five categories: 17-15 optimal; 14-12 good; 11-9 fair; 8-6 borderline; <6 insufficient

Number of positive agreements of the two referees	Number of negative agreements of the two referees								
	0	1	2	3	4	5	6	7	8
8	17								
7	16	15							
6	15	14	13						
5	14	13	12	11					
4	13	12	11	10	09				
3	12	11	10	09	08	07			
2	11	10	09	08	07	06	05		
1	10	09	08	07	06	05	04	03	
0	09	08	07	06	05	04	03	02	01

economist and employed by the university agreed with the answers to each of the five papers, except for one point (only one score "0"). It was not expected that the study design would create the most serious problem. The referees agreed in all five papers that the ideal study design was different from the published study design. In three out of five studies the masking (blinding) of the patients or the intent-to-treat principle created a problem. The length of the follow-up period was too short or the inclusion of all study patients in the evaluation was incomplete in two out of the five studies, and finally, the profiles of the compared study populations or the concealment created a problem in two studies. The statistics were adequate in all five studies.

In summary, this result suggested that the design of the trials rather than the statistics may be problematic. Four out of five randomized controlled trials received a score equal to or more than 9, corresponding to a fair validity. The referees marked one of the studies "borderline" with the score of 8 points.

The students who had just completed a 1-week course on Evidence-Based Health Care used the same procedure as the referees for the critical appraisal of the five publications and came to the following conclusions.

- A Controlled Study of Postoperative Radiotherapy for Patients with Completely Resected Nonsmall Cell Lung Carcinoma by Dautzenberg et al.

As the protocol was changed during the study, neither the original study design nor the power calculation was still be appropriate to answer the study question.

As the patients were aware of the assignment to the treated or untreated group, the result of the study was influenced by this information. The baseline risks were not balanced. The higher risk in the radiotherapy group may explain the higher mortality in this group. If the patients of the control group did not really understand the informed consent, a serious ethical and methodological problem has to be considered as the results of this study were evaluated

Table 3 Results of the referees' and the students' assessments. The referees' assessment was made by a pharmacist from a pharmaceutical company and the FP. The second assessments were made by 20 medical students. The publication was appraised by eight students (four pro-advocates and four contra-advocates) who had to form a consensus. The results demonstrate that the calculated scores may vary considerably

Question	Dautzenberg et al., Cancer 1999		Mehta et al., Lancet 2001		Heart Protrection Study Collaborative Group, Lancet 2002		Meunier et al N Engl J Med Meunier et al., 2004		Piccard-Gebhardt N Engl J Med 2005	
	Referees	Students	Referees	Students	Referees	Students	Referees	Students	Referees	Students
If potentially valid start out with a score of 9 points	9	9	9	9	9	9	9	9	9	9
Add/deduct 1 point if the two study designs are similar/different	-1	+1	-1	-1	-1	Not valid as protocol cannot be reproduced	-1	+1	-1	+1
Add/deduct 1 point if the risk profiles of the compared study populations are similar/different	+1	-1	+1	+1	-1		+1	+1	+1	+1
Add/deduct 1 point if the plan to allocate patients to the study groups were concealed/known to the therapist	+1	+1	+1	+1	+1		+1	0	-1	+1
Add/deduct 1 point if doctors and patients were/were not continuously blinded during the entire period of the study	-1	-1	+1	+1	-1		+1	+1	-1	-1
Add/deduct 1 point if the follow up was long enough/too short to detect the study endpoints in most of the included patients	+1	+1	-1	0	+1		+1	+1	-1	-1
Add/deduct 1 point if all/not all included patients were included in the calculation of the study results	+1	+1	-1	+1	+1		-1	-1	+1	-1
Add/deduct 1 point if all/not all included patients were evaluated according to the intent-to-treat principle	-1	-1	-1	-1	+1		-1	-1	0	+1
Add/deduct 1 point if adequate/inadequate statistics were applied	+1	+1	+1	+1	+1		+1	+1	+1	+1
Total score	**11**	**11**	**9**	**12**	**11**	**<6**	**11**	**12**	**8**	**11**

according to the intent-to-treat (ITT) principle. The application of the ITT principle is not independent from the patient's informed consent: patients in the untreated group may request the promising treatment and therefore reject the random allocation or treated patients may not tolerate the recommended treatment. Patients who are aware of the experimental character of the offered treatment will more frequently express their preference and have a higher

chance of rejecting the offered treatment than patients who are offered only one option (treatment or not treatment) assuming that they are offered the best available standard. Without clarification of this question, the data cannot be interpreted. This is not only a theoretical consideration, the practical importance has recently been discussed (Ohmann 2000). Some of us were not convinced that all patients were really randomized as patients in worse conditions received the more intensive treatment (radiotherapy).

- Effects of Pretreatment with Clopidogrel and Aspirin Followed by Long-term Therapy in Patients Undergoing Percutaneous Coronary Intervention: the PCI-CURE Study by Metha et al.

We do not agree with the design. The study tried to respond to two completely different questions in one trial. The design is acceptable to test the effects of the first 30 days of treatment, but not the effects of long-term treatment. On page 528 myocardial infarction was selected as one outcome but was not routinely screened for. There will be many undetected cases as well as cases that are difficult to assess as positive or negative. One of the take home messages to many participants of our group was the impression that this treatment was recommended for more than 1 year without presenting supportive data. The appropriate follow-up period depends on the peak of incidence of events, which is not mentioned in the article.

- MRC/BHF Heart Protection Study of Cholesterol Lowering with Simvastatin in 20,536 High-risk Individuals: A Randomised Placebo Controlled Trial by MRC/BHF Heart Protection Study Collaborative Group

We could well understand the practical but not really the scientific aim of this study. The rules for selecting and managing the patients were too complicated to use the same rules in a confirmatory study. Patients who responded to placebo were excluded from the trial as well as patients who presented no response to simvastatin. This shows a higher response rate than would be observed in real-world conditions (Knipschild 1991). After an intensive discussion we concluded that the result generated in this study is not relevant for decisions in day-to-day practice.

- The Effects of Strontium Ranelate on the Risk of Vertebral Fracture in Women with Postmenopausal Osteoporosis by Meunier et al.

It is not mentioned whether or not the randomization plan was concealed. No objective methods were described to assess vertebral fractures. The statistical analysis should not be made by the supporting agency. The use of six different tests enhances the risk to select retrospectively a positive result just by chance. The "run-in-phase" or "qualification period" was included. Knipschild et al. (1991) described a list of almost ten effects that can be achieved by introducing this qualification period. The problem is that such instruments will be abused unless the induced effects are carefully appraised.

- Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer by Piccart-Gebhardt et al.

Although most relapses at this stage of the disease are expected after 18 to 24 months, the median follow-up was only 12 month. There is a considerable risk that many patients in the untreated group were lost. The corresponding information is missing. It is a serious flaw not to report the results of the third arm even though these patients had so far received the same treatment as the patients in the second arm. One possible reason not to report these results may a difference between the expected and the observed results. It is surprising that even in a paper published 1 year later, no results of the third arm were reported.

As compared to the referees, the students tended to give slightly higher (better) marks with the exception of the Heart Protection Study. The students thought that this study was seriously flawed by non-reproducible selection and attrition of patients and can therefore not be condisidered as valid information.

Discussion

The systematic assessment of the validity of clinical trials described in this paper has been tested over several years in our student courses. The formal use of the improved method in five publications demonstrates that even so-called level I/II evidence is not free from bias and mistakes. The procedure is easy to apply and seems to be more effective in detecting potential bias than existing instruments (Forestier 2005).

The decision to accept or reject scientific information for policy-making is usually dichotomous: a paper will either be included in a meta-analysis or in a guideline or not. Most scientists and journal editors are convinced that their journals with a high impact factor are more likely to publish level-I or -II articles than journals with low impact factors (Obremskey 2005). This is actually true, but it is also true that bias in journals with high impact factors is more difficult to detect than bias and mistakes in journals with low impact factors (Delgado Rodriguez 2001).

We hypothesized that the TACT scoring system provides the possibility of "fine tuning" the validity of a clinical trial, giving information that may directly influence clinical decisions (Fig. 1). This figure describes that three criteria should be considered when making clinical decisions: the described efficiacy or effectiveness of the compared interventions, the validity of the scientific report that

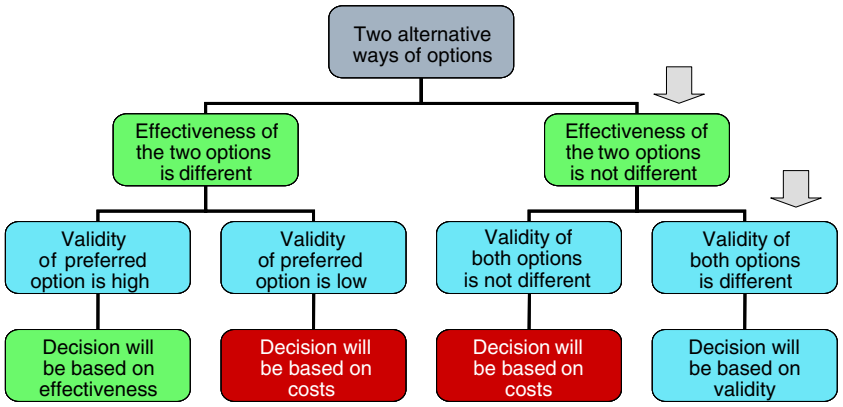

Fig. 1 Decision aid for users of scientific information. Clinical decisions should depend not only on the effectiveness and on costs of an intervention, but should also consider the validity of the data (translated from Porzsolt 2003b)

describes the comparison of the two interventions and the costs of the compared interventions. The most difficult problem is the critical appraisal of the validity of the report. To underline the significance of the validity we included in our study only papers from journals with a high impact factor. The critical appraisal demonstrated that even these papers are not free from bias.

From our EBHC courses we know that the assessment of validity may vary considerably among different referees especially when they come from different groups of stakeholders. Doctors and colleagues from industry will rather play the roles of pro-advocates, while statisticians and epidemiologists will rather take the role of the contra-advocate. These differences in ratings are often based on different expectations and opinions and will arise whenever decisions have to be made. In order to reach such decisions–especially in the highly emotional field of health care–compromises will be inevitable. With increasing complexity these compromises become more difficult. Therefore, it was our goal to design a system for assessment of validity of scientific papers that catalyzes compromises at a very basic stage where rather simple questions can be discussed. The details of the validation process were based on experience with many sophisticated dodges, gimmicks and ploys that were introduced by innovative scientists to avoid random effects. Unfortunately, these tricks frequently induce bias. The only reason to propose our twin assessment of clinical trials is to increase the chance that users of the information can recognize bias.

Even recognized tools, such as the checklist based on the CONSORT statement (Mills 2004), can only identify but not exclude bias. The necessary generation of a formal consensus score seems to be a possibility to disclose controversial issues of referees with different backgrounds and different goals. We demonstrated that two referees with different expectations and goals may provide a combined judgment. Other checklists like the CONSORT statement (Knipschild 1991) and the GRADE system (GRADE working group 2004) are more detailed than our short system, but even these do not address the problems that have to be solved for assessment of the validity when referees with different perspectives are involved.

Conclusion

The formal assessment of the validity of clinical trials using structured techniques and the subsequent appraisal of the assessed results is one side of the medal. The other side refers to the reproducibility of the appraisal and of the final decisions.

The results of our study raise the question on the reliability of structured appraisals by single referees. Single referees or readers of scientific articles have neither the time required for a critical analysis nor do they usually have the necessary experience nor the patience to do this work. Our experiments suggest that group discussions are more effective than assessments by single persons in detecting flaws and bias of clinical trials. In our experiment two referees considered the quality of the Heart Protection Study as fair, while the group of eight students detected too many weak points to consider this study as valid (Table 3). We therefore favor the idea of establishing groups of referees—students or researchers—rather than single experts who are trained, motivated and experienced enough to offer this service. A high quality appraisal may require 2-6 h work for a discussion group.

The problem of variability of critical appraisals can be solved by data that confirm that the results generated under ideal but artificial conditions of clinical trials can be reproduced under real-world-conditions. Until these data are available, decision makers (academic teachers, policy makers) may use external services that provide the information derived from the described critical appraisals.

Acknowledgement We are grateful to Eva Greimel (Department of Gynecology, Medical University of Graz), Martin Eisemann (Department of Medical Psychology, University of Tromsø) and Jörg Sigle (Clinical Economics, University of Ulm) for their contributions to the meeting at Kohldorf/Steiermark, Austria, where the pros and cons of disclosing the details of the publications were discussed.

Conflict of interest Franz Porzsolt is a consultant for Sevier Deutschland.

References

Barry MJ (2009) Screening for prostate cancer—The controversy that refuses to die. New Engl J Med 360:1351–1354

Coppin C, Porzsolt F (2003) Kidney Cancer. Evidence-Based Oncology. In: William C (ed) Evidence-Based Oncology. BMJ Books 333–345

Darmoni SJ, Haug MC, Lukacs B, Biossel JP (2001) Quality of health information about depression on internet. Level of evidence should be the gold standard. BMJ 322:1367

Dautzenberg B, Arriagada R, Chammard AB, Jarema A, Mezzetti M, Mattson K, Lagrange JL, Le Pechoux C, Lebeau B, Chastang C (1999) A controlled study of postoperative radiotherapy for patients with completely resected nonsmall cell lung carcinoma. Groupe d'Etude et de Traitement des Cancers Bronchiques. Cancer 86:265–273

Delgado Rodriguez M, Ruiz-Canela M, de Irala-Estevez J, Llorca Diaz J (2001) Martinez-Gonzalez MA (2001) Differences in the quality of Spanish clinical trials published in international periodicals and of the ones presented in general medicine periodicals with wide readership. Rev Clin Esp 201:437–443

Forestier R, Francon A, Graber-Duvernay B (2005) Validity parameters of clinical trial and their influence on evidence based medicine conception: a review. Ann Readapt Med Phys 48:250–258

GRADE Working Group (2004) Grading quality of evidence and strength of recommendations. BMJ 328:1490–1494. doi:10.1136/bmj.328.7454.1490

Heart Protection Study Collaborative Group (2002) MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet. 360:7–22

Jørgensen A, Hilden J, Gøtzsche PC (2006) Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ 333:782–785

Knipschild P, Leffers P, Feinstein AR (1991) The qualification period. J Clin Epidemiol 44:461–464

Mehta SR, Yusuf S, Peters RJG, Bertrand ME, Lewis BS, Natarajan MK, Malmberg K, Rupprecht H-J, Zhao F, Chrolavicius S, Copland I, Fox KAA (2001) Effects of pretreatment with clopidogrel and aspirin followed by long-term therapy in patients undergoing percutaneous coronary intervention: the PCI-CURE study. The Lancet 358:527–533

Meunier PJ, Roux C, Seeman E, Ortolani S, Badurski JE, Spector TD, Cannata J, Balogh A, Lemmel E-M, Pors-Nielsen S, Rizzoli R, Genant HK, Reginster J-Y (2004) The Effects of Strontium Ranelate on the Risk of Vertebral Fracture in Women with Postmenopausal Osteoporosis N Engl J 350:459–468

Miettinen OS, Henschke CI (2001) CT screening for lung cancer: coping with nihilistic recommendations. Radiology 221:592–596

Mills E, Loke YK, Wu P, Montori VM, Perri D, Moher D, Guyatt G (2004) Determining the reporting quality of RCTs in clinical pharmacology. Br J Clin Pharmacol. 58:61–65 and 58:102

Nothardt (2007) Validity of Clinical Trials in Homeopathy included in Systematic Reviews. Thesis Medical Faculty University of Ulm, 2007

Obremskey WT, Pappas N, Attallah-Wasif E, Tornetta P 3 rd, Bhandari M (2005) Level of evidence in orthopaedic journals. J Bone Joint Surg Am 87:2632–2638

Ohmann C, Albrecht J. Lessons to be learned for gastroenterology from recent issues in clinical trial methodology (2000) Can J Gastroenterol 14: 293–298

Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M. Levels of evidence. Homepage des Oxford-Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?0=1025

Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, Gianni L, Baselga J, Bell R, Jackisch C, Cameron D, Dowsett M, Barrios CH, Steger G, Huang CS, Andersson M, Inbar M, Lichinitser M, Lang I, Nitz U, Iwata H, Thomssen C, Lohrisch C, Suter TM, Ruschoff J, Suto T, Greatorex V, Ward C, Straehle C, McFadden E, Dolci MS, Gelber RD; Herceptin Adjuvant (HERA) Trial Study Team (2005) Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med 353:1659–1672

Porzsolt F, Kumpf J, Coppin C, Pöppel E (2003a) Stringent application of epidemiologic criteria changes the interpretation of the effects of immunotherapy in advanced renal cell cancer. In: William C (ed): Evidence-Based Oncology. BMJ Books 34–38

Porzsolt F (2003b) Klinische Ökonomik. Die ökonomische Bewertung von Gesundheitsleistungen aus der Sicht des patienten. In: Porzsolt F, Williams AR, Kaplan RM (Hrsg.) Klinische Ökonomik. Effektivität and Effizienz von Gesundheitsleistungen. ecomed Verlagsgesellschaft 17–40

Porzsolt F, Kajnar H, Awa A, Fässler M, Herzberger B (2005) Validity of original studies in Health-Technology Assessment (HTA) reports: Significance of standardized assessment and reporting. Int J Technol Assess Health Care 21:1–4

Straus SE, Richardson WS, Glasziou P, Haynes RB (2005) Evidence-Based Medicine. How to Practice and Teach EBM. 3rd edition. Elsevier Churchill Livingstone, Edinburgh, London New York

van Gijn J (1999) From therapeutic trials to current practice. Rev Neurol (Paris) 155:708–712